The Journal of
HISTORICAL
NETWORK
RESEARCH

12 | 2025
special issue

Time and Temporality as Categories of Connectedness

CINDARELLA PETZ | ALINE DEICKE | JULIAN JAROSCH

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

uni.lu
UNIVERSITY OF
LUXEMBOURG

The Journal of
HISTORICAL
NETWORK
RESEARCH

MALTE VOGL/ROBERTO LALLI

Temporalities in the spread of knowledge

Journal of Historical Network Research 12 (2025)

Keywords socio-epistemic networks, co-citation networks, time-dependent clustering, history of science

Abstract Socio-epistemic networks are an analytical and methodological framework that has been developed to explore the interplay of the social, material and semantic dimensions of the structural changes in knowledge systems by using the concepts and tools of network science. This framework is suited to analyzing the development of scientific fields throughout the 20th century, which traditionally presents several challenges: the scales of change are often gradual; the definitions and boundaries of these fields are typically constructed retrospectively; and the perspectives on the structure of a field at any given time can vary significantly among key actors. Additionally, traditional scientometric techniques used for historical analysis have major limitations, as they often rely on presentist assumptions regarding the metadata format of available online publications repositories.

In this methodological contribution, we aim to overcome these limitations by proposing a workflow that uses temporal co-citation clustering of publication corpora. This workflow is part of an iterative process involving cooperation between digital humanists and historical domain experts. The goal is to extract meaningful, temporally extended clusters of publications that represent evolving scientific fields. We utilize the Leiden community detection algorithm to delineate these clusters. For the time scales of evolution of each retrieved cluster, further analytical steps are conducted to examine the temporal shifts in centrality measures of specific institutions or actors in the field, thereby revealing their changing roles within the research field. The approach is exemplified through a case study on the evolution of astrophysics and related research fields, and the role that the Max Planck Society played in these dynamics.

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITY OF
LUXEMBOURG

Introduction

The quantitative study of the development of scientific fields throughout the 20th century usually relies on metadata retrieved from online platforms. These data are collected by third-party institutions, often for-profit companies, with their own collection logic, which does not have historical research as its primary scope.[1] As a consequence, they offer ahistorical views based on contemporary categorizations of research that might not necessarily apply to the past.[2] In some cases, the categorization is based solely on the venues in which specific articles are published, providing little granularity and precision in the identification of scientific fields.[3] Some historical analyses address these challenges by analyzing a preselection of specific journals or scientific organizations.[4] Other researchers start with a keyword search to define the corpus related to a research field.[5] These approaches, however, necessitate either a foundational knowledge of the field's evolution from the outset of the analysis, or limit the validity of the findings to specific sectors within the field.

To overcome these limitations in the historical analysis of recent science, in this methodological contribution we introduce a workflow that makes use of only the reference information included in publication records, and iteratively creates and analyzes sub-corpora in relation to a given research question. Our approach is based on the idea that clustering analysis of co-citation networks represents research fields and their developments over time. While the analysis of co-citation networks as representations of research fields is widely employed,[6] its application for historical analysis is much less developed. These studies typically compare

Corresponding author: Malte Vogl, Department Structural Changes of the Technosphere, Max Planck Institute of Geoanthropology, Jena, Germany vogl@gea.mpg.de
Acknowledgements: The authors acknowledge fruitful discussions with the ModelSEN team: Aleksandra Kaye, Raphael Schlattmann, Jascha M. Schmitz. Lea Weiß, Laura von Welczeck and Dirk Wintergrün. This research has made use of NASA's Astrophysics Data System Bibliographic Services. Part of the algorithm for generating the reports is based on work by Irina Kokoshko.
Funding Information: The research project ModelSEN was funded by the Federal Ministry of Education and Research, Germany (Grant No. 01 UG2131).
Data accessibility: The python code for the workflow of this paper is published as part of the package SemanticLayerTools on PyPI: https://pypi.org/project/semanticlayertools (last viewed June 5th, 2024) and on Zenodo https://doi.org/10.5281/zenodo.10281812. The documentation is available at https://semanticlayertools.readthedocs.io (last viewed June 5th, 2024). The publication of ADS data is not possible due to copyright restrictions.

1 Harzing 2019.
2 Bornmann 2018.
3 Perianes-Rodriguez and Ruiz-Castillo 2017.
4 See, e.g., Wagner and Leydesdorff 2003.
5 See, e.g., Bettencourt et al. 2008.
6 Boyack and Klavans 2010; Klavans and Boyack 2017.

clusters of co-cited papers in different periods, based on the publication years of the citing papers. Even in more elaborate approaches that differentiate between intellectual bases (cited papers) and research fronts (citing papers), like those implemented in the software *Citespace,*[7] researchers can only retrieve clusters limited by the specific fragmentation of the historical period selected at the start of the analysis.

Our proposed methodology makes use of the Leiden community detection algorithm[8] to disentangle scientific fields from large corpora and explore their evolution over a long period of time, with no preliminary decision for restricting the corpus at the beginning of the analysis. This paper presents a detailed description of the workflow implementing this methodology for a specific case study of historical research.

In the following subsections of the introduction, we first briefly illustrate the framework of socio-epistemic networks, which constitutes the conceptual background of the analysis. Subsequently, we discuss different analytic approaches that become possible within this framework, including a reflection on the concepts of temporality that underlie different approaches. We then illustrate the chosen case study on the evolution of astrophysics and astronomy, concluding with a brief description of the publication dataset chosen to investigate this case study.

Conceptual framework: Socio-epistemic networks

The conceptual framework for this contribution originates in research undertaken by Jürgen Renn and his colleagues at Dept. I of the Max Planck Institute for the History of Science. Renn and his colleagues define knowledge as a "codified experience" with the "potential for problem-solving."[9] This codification can be interpreted as the interplay of social, material and cognitive structures of knowledge production and circulation, as all activities related to knowledge production must be performed by a group of actors. Examples include agreeing on a sign system or an institutional form of social interactions, as well as the production of material realizations of knowledge, such as scientific publications or technical instruments. Knowledge is therefore seen as a dynamic and multi-faceted process with three main interlinked dimensions of analysis: social, material and semantic.

---

7 Chen and Song 2018.
8 Traag et al. 2019.
9 Renn et al. 2016, pp. 35–79, 37 (translation ours).

This three-part categorization of knowledge evolution, called socio-epistemic networks, inspired the project Modeling Socio-Epistemic Networks (ModelSEN)[10] funded by the Bundesministerium für Bildung und Forschung-BMBF (Grant No. 01 UG2131). The project aims to develop methods and routines for applying the concepts and tools of network theory to the investigation of the historical developments of knowledge systems.[11] The project is based on the fundamental assumption that knowledge development can be modeled as a multilayered, evolving network of social, material, and semantic nodes connected by various types of edges.[12] Although each layer of the socio-epistemic networks depends on distinct data sources and requires specific routines for creating nodes and edges, they are not strictly separable.

An example of such a multilayered structure is the set of publications of a chosen scientific research institution. Each paper is a material representation of knowledge, created by its authors, who use specific language to convey semantic concepts. Authors are connected to other authors by social connections, e.g., by their co-authorship of a paper, retrievable in the publication data. Authors are connected to the publications they have written. Publications can also be connected in different ways, e.g., by appearing in the same special issue, or by being co-cited together at a later point in time. The groups of words employed in the publication, or "n-grams", are connected to the publication itself, but also to each other, e.g., by a similarity measure.[13]

Levels of analysis

The proposed multi-layered analysis allows the exploration of a chosen publication dataset at different scales. By creating and analyzing socio-epistemic networks of large corpora, researchers can uncover large-scale structures of past dynamics, that is, a macroscopic view of research fields' development. To this macroscopic view, a more microscopic view can be added. This might focus, for instance, on the position and role of a single institution within the socio-epistemic network, addressing specific questions such as the effectiveness of its funding structure on scientific output or the development of external collaborations by that institution.

In addition to considering different scales, the modeling of socio-epistemic networks must also take into account the varying perspectives on temporality

10 https://modelsen.gea.mpg.de (last viewed June 6th, 2024).
11 For applications of this approach to the history of general relativity, see Lalli et al 2020a; Lalli et al 2020b.
12 Kaye et al. 2024.
13 Bizzoni et al 2021.

that can be employed. Analyzing the structure of a given corpus often reflects a 'post-facto' perspective on the past, which can potentially obscure the complex motivations behind knowledge creation. A more granular view may reveal potential cause-and-effect relationships, such as the impact of governmental funding decisions on scientific outputs for, e.g., a single country or scientific institution. An analysis of cause-and-effect relationships must consider additional source material, e.g., meeting protocols of science funding organizations, which makes an automated analysis much more challenging. On an even more detailed level, each researcher generates knowledge within a personal temporal frame, having access to only a fragment of past knowledge and lacking foresight into the future relevance of their work. The effect of such personal temporalities on the development of scientific thought (and thus publications) could, for example, be uncovered by studying personal notebooks of researchers. However, such corpora are most likely preserved only for more famous scientists and in hand-written form. An analysis of the evolution of scientific thought at this level of granularity aligns more closely with traditional historical methodology than with quantitative analysis.

In this contribution, we demonstrate how the proposed methodology can address the various scales and temporal dimensions of the material layers of knowledge evolution by applying it to a case study on the history of astrophysics and astronomy, which emerged within the framework of a different historical project.

Case study: The role of the Max Planck Society in the development of astrophysical and astronomical sciences

The Research Program on the History of the Max Planck Society (GMPG)[14] has investigated the multiple facets of research at the Max Planck Society for the Advancement of Science (MPG, after its German name) from the perspectives of contemporary political, social and economic history, as well as of the history of science.[15] The MPG is one of the main German research institutions officially dedicated to promoting and pursuing frontier investigations in basic sciences. The timeframe of the GPMG project goes from 1948 – when the MPG acquired its current name within the post-World War II reconfiguration of basic research in what was to become West Germany – through to 2004. In this period, about 100 MPG institutes have operated in different research domains, spanning the humanities and social sciences as well as, more extensively, STEM disciplinary domains. In order to deal with such a vastity of research activities spanning almost sixty years, the project has elaborated an analytical framework that captures the research structure of the MPG, grouping together various researchers and institutes. This

14 https://gmpg.mpiwg-berlin.mpg.de (last viewed June 6th, 2024).
15 Renn et al., (ed.), 2024.

key analytical framework posits that clusters of research fields emerged within the MPG, shaping future developments of its research portfolio. These clusters are characterized by networks of scientists who are connected either by working in related disciplinary areas, or by utilizing similar scientific tools, or again by participating in the same self-governance structures within the MPG.[16]

Using this approach, the fields of astrophysics, astronomy, cosmology and space sciences have emerged as an especially large and influential cluster within the MPG. We shall call this the *astro cluster* for short. Bonolis and Leon have investigated the development of this cluster within the MPG by making use of the traditional hermeneutics of historical research.[17] In their extensive work, the authors have both provided a periodization of research in all sciences related to the astro cluster and identified the major trends and roles characterizing research in these fields of MPG institutes.

To evaluate the evolution of this cluster from a more quantitative perspective, additional research has been conducted within the ModelSEN project. The primary goals were to quantitatively map the development of research areas related to the astro cluster within the global publication landscape from 1948 to 2004 – the timespan of the GMPG project – and to investigate the specific role and influence of the MPG in this global context. In what follows we describe the step-by-step workflow developed to accomplish this task, serving as an example of a general methodology aimed at analyzing the development of scientific fields in the 20th century.

The dataset: Astrophysics Data System

To pursue our quantitative analysis of sciences related to the astro cluster, we chose the database operated by the Smithsonian Astrophysical Observatory (SAO), called SAO/NASA Astrophysics Data System (hereafter ADS).[18] While the ADS is a digital library portal specifically dedicated to researchers working in fields related to the above-mentioned astro cluster, it contains publications in "astronomy and astrophysics, physics, and general science," thereby considerably broadening the scope of the repository.[19] During the initial exploration of the ADS corpus, we confirmed that the database contains many more papers than initially expected on disciplines unrelated to the astro cluster. Not only are there many publications dealing with physical sciences in general, such as solid-state or particle physics, which could, in principle, albeit loosely, be related to the astro

16 Reinhardt 2024, p. 224.
17 Bonolis and Leon 2022. See also, Bonolis and Leon 2020, pp. 285–361.
18 https://ui.adsabs.harvard.edu/ (last viewed June 4th, 2024).
19 https://ui.adsabs.harvard.edu/about/ (last viewed June 13th, 2024).

cluster, but it also indexes publications clearly out of scope of our research focus (e.g., publications in behavioral or educational sciences).

Since the ADS does not provide any research field classifications of the corpus, an iterative workflow to disentangle the different research fields contained in the corpus was developed. In the next section, we present the workflow for data access and preparation.

Data access and preparation

The ADS search engine permits specific queries for author, year, etc. However, for the scientometric analysis undertaken in the project a copy was created using SOLR on local servers to allow faster access. This copy was created based on an ADS data snapshot from August 2021; later corrections on the live dataset are therefore not reflected in this analysis. To cover the research period, a query was made to the local SOLR endpoint for each time range using the pysolr package and each result set was loaded as a Pandas DataFrame. In what follows we describe the workflow for the creation of yearly co-citation graphs and for the exploration of the datasets and co-citation networks.

Extrapolation and analysis of institutions' data

To evaluate the involvement of the MPG and compare it with other leading research institutions, three regular expressions were compiled to identify and extract affiliation data for three major organizations with significant programs in the astro cluster research fields: the MPG, the University of California System (UCS), and Harvard. After checking the DataFrame affiliations entries for matches, the temporal evolution of the publication counts for these institutions was plotted in relation to the overall publication output for each year from 1950 to 2004 (Fig. 1). This timeframe slightly differs from the above-mentioned research period of the GMPG project, since publication records before 1950 did not consistently contain affiliation data of the MPG. The publication trends of the broader UCS universities closely mirrored the overall publication dynamics, whereas the MPG exhibited a more distinct behavior. Prior to 1970, the publication output of the MPG was comparable to that of Harvard. However, post-1970, the dynamics of the two institutions diverged, potentially due to the increased number of Max Planck Institutes. Such metadata on institutional involvement will be utilized in a subsequent analytical step.

The resulting datasets for each year between 1950 and 2004 were saved in JSON file format to preserve the list structure of the references of a publication.

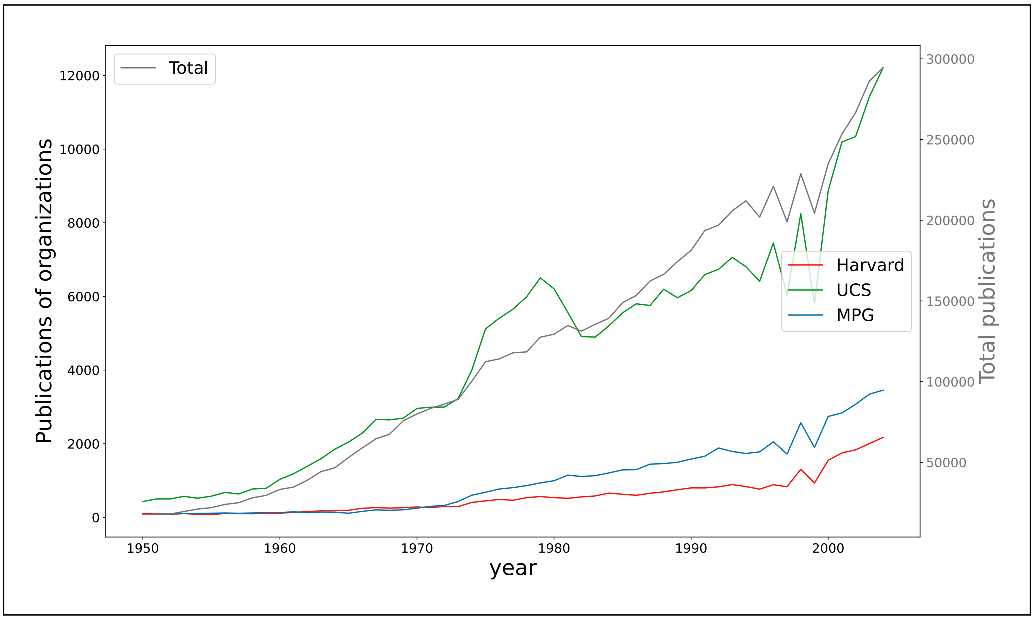

Fig. 1 Publications by the MPG (blue), the University of California System universities (UCS) (green), and Harvard University (red) compared to the overall publication number in the ADS corpus, 1950–2004. The total number of publications in the ADS and the number of publications by institutions are plotted on two different scales on the y-axes.

Creating co-citation edges

In the subsequent preparatory step, for each publication entry within the entire 1950–2004 ADS corpus in each year's DataFrame a weighted list of co-citation edges was written to an edge file. For a given year, all co-citation edges were collected and counted for repeating pairs of referenced publications. This creates the weight of the co-citation edges. This process is computationally intensive, particularly for papers with a large number of references, such as review articles. For instance, a single paper with 100 references can result in approximately 4,900 co-citation edges. Throughout the entire timespan of our corpus, the proportion of extensively referenced papers – which we consider as publications with more than 100 references – ranges between 0.1% and 0.3% of the total corpus. Such a small fraction could justify their exclusion from the study to make the analysis feasible. However, by employing parallelization strategies on a large workstation, we were eventually able to include all publications in the co-citation edge creation process.

Using the Python package igraph, a yearly graph was created from this list of edges, and the giant component (GC) for that year was calculated and saved in Pajek format. Moreover, the number of nodes and edges of the GC were compared

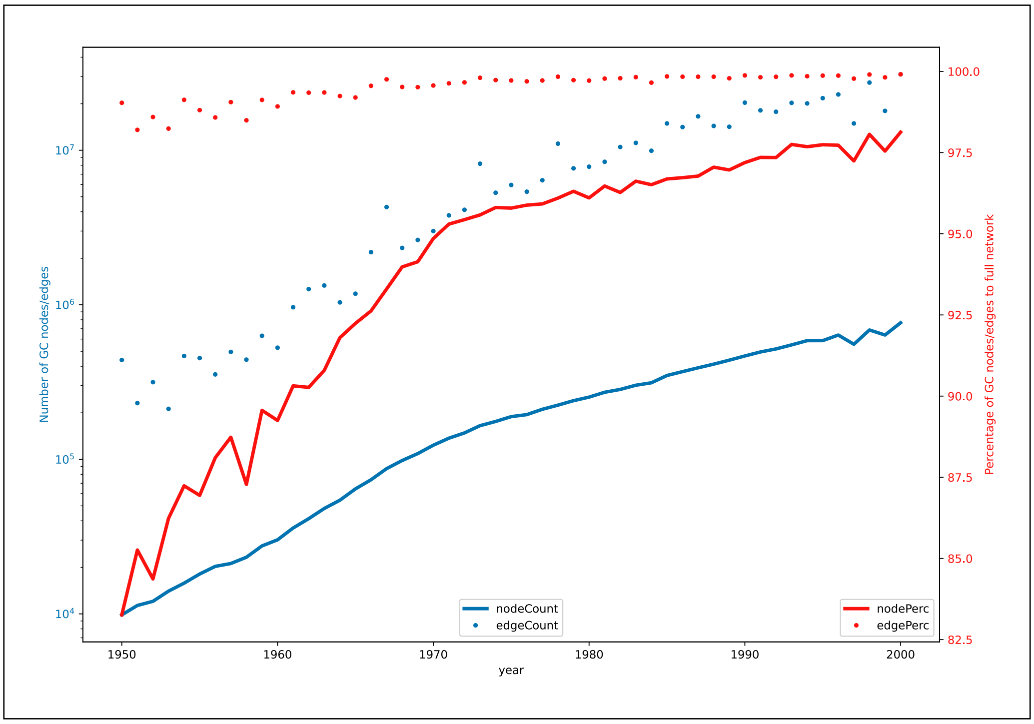

Fig. 2 The number of nodes (solid blue) and edges (dotted blue) as well as the percentage of nodes (solid red) and edges (dotted red) within the giant component (GC), compared to the total nodes and edges for each year. Note the use of two different y-axes, with values displayed on a logarithmic scale.

to the total number of nodes and edges of each yearly graph (Fig. 2). The diagram shows that the number of cited publications increased by nearly two orders of magnitude between 1950 and 2000. Fig. 2 also illustrates that the percentage of the nodes in the GC relative to the total number of nodes steadily increased over the years, showing two different regimes before and after 1970. By contrast, the vast majority of edges in the yearly graphs of co-citation networks belonged to the GC throughout the entire period, with the percentage exceeding 97 % shortly after 1950.

The last exploratory step was to compare the time difference from the source of a co-citation pair to its references, as two co-cited publications might be from vastly different years in the past as seen from the citing publication. To do so, we can make use of the ID structure of ADS publications, where each unique ID starts with the publication's year. For a given pair of co-cited publications, we extract the source and target paper of the edge and multiply each with the weight of that edge. We then count in each year the weighted number of all referenced years and normalize that number by the total number of references found. To enable a comparison across different years of co-citation, we can calculate the

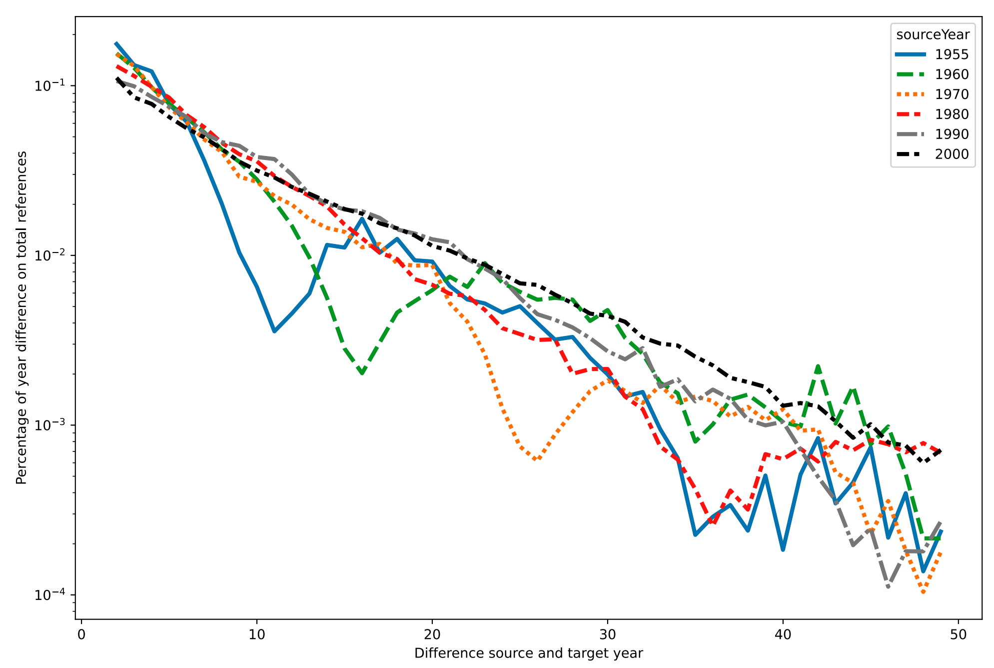

Fig. 3 Percentage of relative age of references in relation to the year of the citing publications in ADS for the years 1955, 1960, 1970, 1980, 1990, and 2000. The diagrams show the percentage difference between the years of the source (the citing publication) and the target (the cited publication) for references co-cited at least once, i.e. only considering sources with at least two references.

relative temporal difference between referencing year and referenced year. For example, if a paper published in 1960 cites two publications from 1957 and 1955, it will give rise to references with ages of 3 and 5 years, respectively. Fig. 3 shows the temporal evolution of these relations for a range of years of the citing papers between 1955 and 2000. The logarithmic scale of the y-axis shows the exponential 'forgetting' of older publications as seen by citation patterns. This feature appears relatively stable for the past 50 years, only disrupted by major events like the Second World War, visible as a dip in the graph, e.g., with a 10-year difference for the 1955 graph (solid blue line), corresponding to an increased level of forgetting older research.

The co-citation networks created consist of millions of edges and a number of nodes between ten and a hundred thousand, even when the analysis is limited to their giant component parts. Attempts at visualizing such large networks, even for a single year, result in nothing more than 'fuzzy hairballs'. A comparison across several years thus necessitates a structural analysis, which will be presented in the following.

Temporal clustering of citation networks

While the clustering of networks is a well-researched problem,[20] the analysis of clusters in longitudinal networks is a relatively new research field, as is the analysis of clusters in multilayer networks. For the workflow of cleaning the ADS corpus and finding the research fields contained within, we made use of the package leidenalg by Vincent Traag.[21] We then used the basic idea of Mucha et al. to consider time as an additional layer in the network and create inter-layer edges between adjacent year layers.[22] For the nodes, i.e., the co-cited publications, the edges between year layers are created between nodes that share the same identification (e.g., the same DOI). The hierarchical clustering algorithm then iteratively uses a quality function to determine whether a node should be part of a cluster or not until an optimal solution of clusters is determined. This process can be either limited by a given number of iterations to run, or be set to stop after new iterations lead to no changes in the mapping between nodes and clusters.

The quality function *Constant-Potts-Model* was chosen for this analysis since it is well defined for weighted graphs and allows finding communities of different sizes, depending on the chosen parameters.[23] Its mathematical structure is given by

$$Q = \sum_c^m \left[m_c - \gamma \binom{n_c}{2} \right]$$

where m_c is the number of internal edges in cluster c, n_c is the number of nodes in cluster c and γ is the resolution parameter.[24] The key parameter here is the resolution parameter γ, which determines the internal density of the found clusters. Any found cluster needs to have a density at least as high as the resolution parameter. The density between two different clusters needs to be lower than the resolution parameter. This also allows fine-tuning the retrieved clusters to the research question at hand. For higher resolutions, one might find clusters of publications for a group of institutes, but which neglects larger scale structures. For finer resolutions, larger scale structures emerge at the cost of smoothing out the finer collaboration pattern.

An exploratory initial phase was required to find a suitable parameter set for the clustering in relation to the historical research question. This can require a large amount of computing time, depending on the source dataset. For the GC of the co-citation network derived from the ADS dataset spanning 1950 to 2004, the

20 Lancichinetti and Fortunato 2009.
21 Traag et al. 2024.
22 Mucha et al. 2010.
23 Traag et al 2011.
24 See e.g., the leidenalg Documentation for details (last viewed June 4th, 2024).

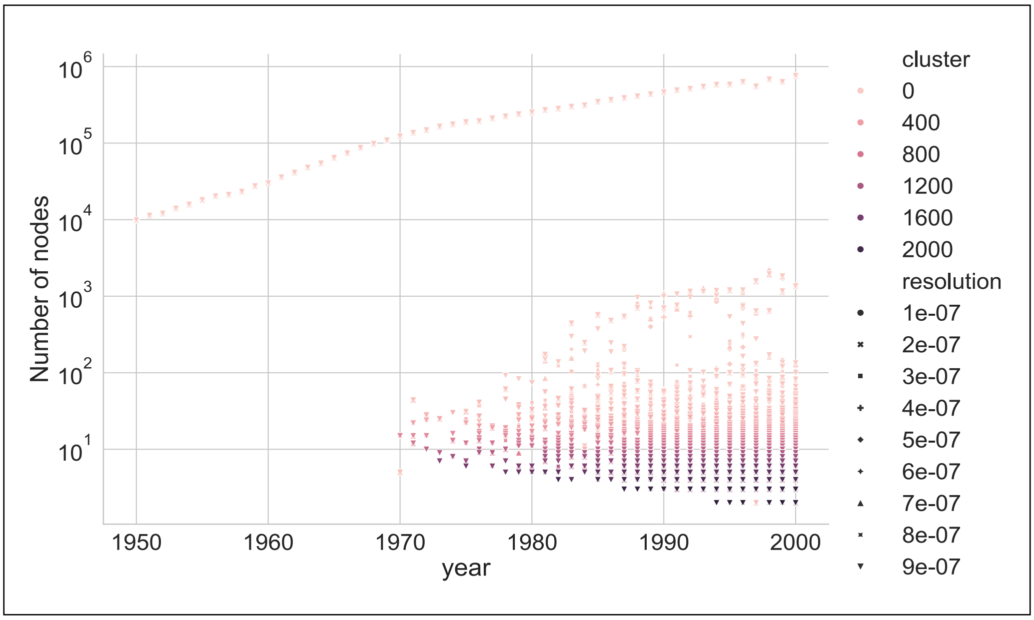

Fig. 4 The relation between resolution parameters and the number and size of retrieved clusters of the co-citation network based on the ADS dataset from 1950 to 2004. The resolution parameter varies from $1 * 10^{-7}$ to $9 * 10^{-7}$. Increasing the resolution leads to finding more mid-sized and small clusters (darker color), while keeping the main large cluster (light upper symbols) stable.

number of nodes totals 12 million, with 430 million edges. These values have to be retained in memory to allow the optimization algorithm of the leidenalg package to find the optimal solution. Running this on a top-of-the-range workstation required more than 400 GB of RAM and eight hours of computing time for a single parameter set. The result of a parameter sweep is shown in Fig. 4, where the resolution parameter was varied; the resulting number and size of clusters is shown.

Depending on the chosen resolution parameter, the retrieved clusters reflect different types of scientific research field structures. These can consist of publications belonging to cooperating institutions for clusters with around 5,000 publications or less, publications related to a specific innovation in a sub-field, e.g., a new type of laser diode, with more than 10,000 publications, or correspond to full (sub-)disciplines, e.g., earth-system science with more than 100,000 publications.

Once a useful set of resulting clusters is retrieved, they are analyzed for internal coherence with respect to aspects such as their scientific field, employed instrumentation or authorship. The parameters employed in the current workflow are listed in Table 1.

Resolution parameter (iteration 1 & 2)	$3 * 10^{-6}$
Interlayer coupling	0.3
minClusterSize	6000

Tab. 1 Workflow parameters to reproduce Fig. 7

The clustering routine adheres closely to the guidelines provided by the author of the Leiden algorithm for clustering in multiplex networks.[25] The graphs are loaded into a dictionary and a convenience routine is used to create layers and interslice layers. The clustering quality function is initialized for each layer and interslice layer. After running the optimization routines, the optimal subgraph partition is returned.

The result of this iteration step is achieved by converting the internal slice notation back to the corresponding year for each time slice. The output of this iteration step is a ledger that includes information about the publication ID, the year it was cited, and the cluster identifier for that year. The same publication can reoccur as part of a co-citation cluster in later years, which can be interpreted as the permanence of a core publication related to the scientific identity of the cluster. Another possibility is the substitution of one publication by another, while keeping the cluster identity stable, which can be interpreted as an update of the core papers that constitute the cluster.

All these processes must be noted in the resulting dataset, leading to a rather large set of about 12 million entries. To get a first impression of such a large dataset for historical interpretations, the workflow allows the creation of streamgraphs. Here, the number of nodes in each cluster is calculated for each year. Subsequently, the values for each cluster are stacked and plotted, with each cluster assigned a fixed color. To get an impression of the temporal evolution of this data, the y-axis is removed and values are grouped around a central abstract y-value.

Figure 5 shows the visualization of a streamgraph of the cluster sizes over the period of our historical research obtained by using the altair software.[26] It appears that while a few clusters cover the full period, others emerge only in later periods. By interpreting clusters of co-cited papers as research fields, the analysis suggests that new scientific fields emerged both around 1970 and around 1990, and that these grew in the following period to such an extent that they came to occupy a large fraction of the entire network at the end of the period of our analy-

25 See section on temporal community detection in the leidenalg multiplex docs. (last viewed June 4th, 2024).

26 Vanderplas et al. 2024.

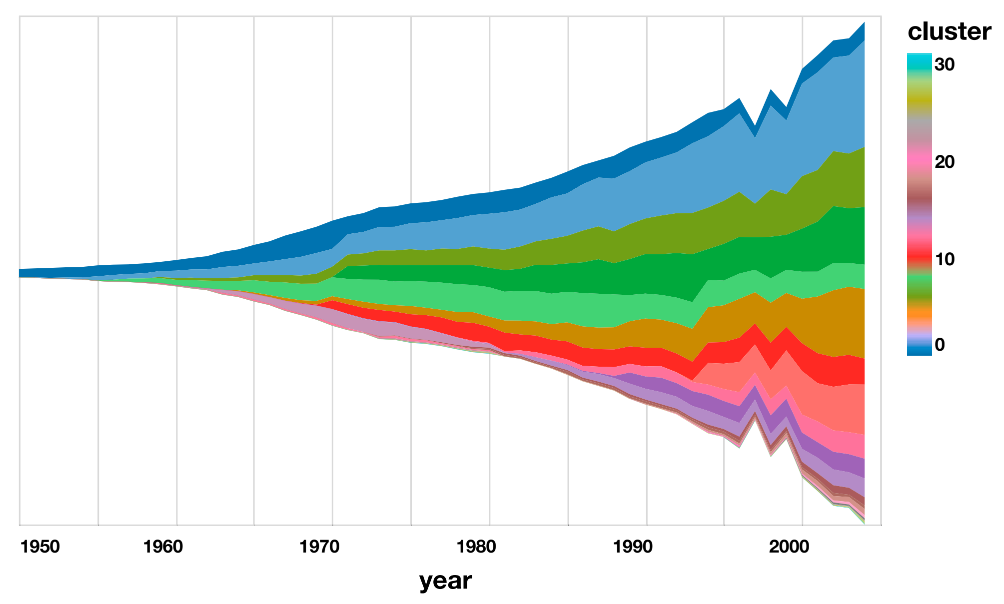

Fig. 5 The streamgraph of the cluster sizes per year of the co-citation network of the ADS corpus after the first iteration. The data for this figure was selected for clusters with more than 600 publications, leading to more than 30 clusters.

sis. An analysis of the clusters' content will be presented in the following section, but it can be anticipated that the temporal effects visualized in the streamgraph appear to represent recognizable historical processes.

Until this point of the description of the workflow, the clustering only made use of information contained in the co-citation patterns. Such information is not readily applicable to historical enquiries. An additional step is necessary to connect structural information with the actual scientific content.

Reports of the clusters' content

The next step of the workflow consists of an analysis of the retrieved clusters to ascertain their scientific meaning. To do so, the metadata on affiliations, authors and titles of the publications in the ADS database are considered. For each cluster with a size greater than a chosen threshold, called minClusterSize (see Table 1), the available data are collected and presented as a report (see Fig. 6), which is then analyzed by the historian of the cooperating group. This report of the cluster contains the following pieces of information: its size, i.e., the number of nodes; the top 20 authors and affiliations, i.e., the 20 scientists and institution names that appear most often in the metadata of the cluster's publications; and, cru-

```
Got 284609 unique publications in the time range: 1962 to 2004.
    Found metadata for 265474 publications.
    There are 19135 publication(s) without metadata.

The top authors of this cluster are:
    Seinfeld, John H.: 251 pubs
    Hobbs, Peter V.: 246 pubs
    Broecker, Wallace S.: 151 pubs
    Fehsenfeld, F. C.: 150 pubs
    Krishnamurti, T. N.: 147 pubs
    ….

The top 20 affiliations of this cluster are:
    NASA, Goddard Space Flight Center, Greenbelt, MD: 835 authors.
    NASA, Langley Research Center, Hampton, VA: 815 authors.
    NASA, Ames Research Center, Moffett Field, CA: 455 authors.
    JPL, Pasadena, CA: 417 authors.
    National Center for Atmospheric Research, Boulder, CO: 390 authors.
    ….

Finished analysis of cluster 7 with 284609 unique publications in 9167 seconds.
```

Fig. 6 Example of a report for a cluster of the co-citation networks built on the ADS datasets (1950–2004)

cially, the timespan of the cluster, defined as the oldest and most recent edge of the co-cited paper included in the cluster, which is the year of publication of the citing paper. It should be noted that metadata collection can only make use of the ADS data already gathered. In the case of co-cited publications that are not themselves indexed in ADS, we cannot retrieve information about the authors and their affiliations. To get an understanding of the availability of the metadata sets, the report also contains the number of documents for which metadata was found and the number of publications that had no metadata.

A topic modeling algorithm is employed based on the title data of the cluster's publications. Using the textacy package[27] and the Nonnegative Matrix Factorization (NMF) algorithm,[28] the analysis considers words which are in at least two documents but are not present in more than 95 % of the corpus. Since this topic

27 DeWilde et al. 2024.
28 Févotte and Idier 2011.

modeling step would again be time-consuming for the complete cluster corpora, the workflow creates a sample corpus of 1% of the given titles for each cluster. Based on this sampled corpus data, the topic modeling is run for first 15 and then 50 topics, to compare the effect of the number of topics on the quality of the topics retrieved. The algorithm returns a listing of the topics with the top 20 words describing the content of each topic.

Based on these reports, one of us (RL) selected clusters of papers which do not fit the scope of the research question, as they show no relation to the themes of astrophysics, astronomy, cosmology or space science. These included, for example, clearly identifiable large clusters of research fields like solid state physics, quantum optics, superconducting physics and earth-system science. This selection is then used to create an exclusion list, which is considered during the next iteration of the workflow.

Historical interpretation of temporal clusters

The remaining clusters, which were considered as potentially part of the astro cluster, were manually tagged with labels corresponding to the found authors, institutions and topics. The selected clusters can then be again represented as a streamgraph, this time with meaningful cluster names (see Fig. 7). A tentative historical interpretation of the macroscopic view represented in the streamgraph could then be produced. At this stage, the historical interpretation is used to evaluate the robustness of the methodology and the degree of information that could be retrieved with the chosen parameters. From the given labels, it appears evident that the degree of granularity obtained after the second iteration using the chosen parameters only enables us to separate the papers into major research areas within the astro cluster.

The report of the black cluster, which spans the entire period of our analysis, indicates that it is a representation of nuclear physics. The fact that it is the only cluster visible in the first decade of our analysis simply indicates that in this period, nuclear physics was the most quantitatively present in the co-citation networks. This outcome may be attributed to the rapid growth and consolidation of nuclear research using accelerators as a big-science field, where an increasing number of papers are frequently cited together to support large-scale scientific projects.[29]

The following period saw the progressive emergence of other fields; by emergence we mean clusters becoming large enough to appear in the visualization,

29 For a seminal volume on the rise of big science underlying the role of nuclear physics in the post-World War II period, see Galison and Hevly 1992.

with the rise of astrophysics and observational astronomy (red) as well as chemical physics and physical chemistry (blue) starting from the late 1950s. Keeping in mind that these clusters are retrieved in the second iteration after some clusters of the initial ADS co-citation network have already been excluded (because they were judged to be unrelated to astro cluster research areas), we might tentatively try to interpret these signals. The growth of astrophysics and observational astronomy in Fig. 7 strictly correlates with the beginning of the space age, with all the implications this had for new technical possibilities for astronomical and astrophysical research, as well as the role of space-related sciences in the Cold War rivalry.[30] Contrary to the red cluster, chemical physics and physical chemistry do not refer only to research that is relevant to astro cluster science, but instead covers a broader disciplinary area. The effect of the emergence of the red cluster during the same period might be related to more general transformations in research funding, novel technologies, industrial applications, and the use of electronic computers to support calculation and modeling.

In the early 1960s, one sees the emergence of geosciences (dark green), which might be partially related to the increasing relevance of this discipline for matters related to Cold War issues, such as surveillance technologies in relation to the many efforts to increase the availability of data starting from the International Geophysical Year, with the creation of World Data Centers.[31] One enigmatic cluster is the mustard colored one, which represents solid-state research lasting only from the early 1960s to the late 1970s. This is difficult to interpret, given the increasing relevance of this research area within the physics landscape.[32] The most probable interpretation is that this cluster is simply a remnant of the exclusion step performed after the first iteration, where a cluster identified as solid-state physics was excluded from the analysis in the second iteration. The resulting small solid-state-physics-related cluster in Fig. 7 is mostly related to research on a subset of solid-state physics focusing on the electronic and structural properties of materials.

In the late 1960s, another cluster that rose to prominence is identifiable as the research area of space physics and solar-stellar interactions with a specific focus on space plasma physics (yellow). The emergence of the field during this period is related to a combination of historical factors: the new avenues that opened for research in the space race, with the first important discovery being the Van Allen radiation belts surrounding the Earth; and the declassification of thermonuclear

30 DeVorkin 2010.
31 Turchetti and Roberts, eds., 2014; on World Data Centers, see Aronova 2017; the ERC AdG project NEWORLD@A: Negotiating World Research Data – A Science Diplomacy Study is currently working to uncover the historical role of World Data Centers in the development of the current system of global circulation of scientific research data.
32 Martin 2019.

fusion research in 1958, which resulted in the publication of a great number of influential papers on theoretical plasma physics by the early 1960s.[33] Keeping in mind that the year is defined by the time the edges are created in the co-citation network, namely by the year of the citing article, Fig. 7 most probably shows the slightly delayed effects of the two major historical processes just outlined.

The 1970s are characterized by the emergence of a field that might be termed atmospheric physics (light green), which grew substantially in the following period. Even in this case it is possible to relate this effect in the visualization with a historical process related to the increase of environmental concerns, which had an important turning point in the early 1970s, as epitomized by the 1972 United Nations conference on Human Environment held in Stockholm.[34]

The big picture of the scientific evolution of research areas related to the astro cluster thus shows a further specialization in the 1980s with new clusters emerging: fluid dynamics (light blue), atomic molecular and optical physics (brown), and solar system science (dark red). While the dark red cluster is clearly part of the spectrum of research fields related to the astro cluster, the other two clusters are relevant to astro-cluster sciences but are broader in scope, encompassing two fields that extend beyond their direct connection to astrophysical research.

The most recent clusters shown in Fig. 7 are the pink cluster that explodes in the 1990s and the gray cluster that emerges in the early 2000s. The pink cluster groups together publications in theoretical physics that delve into extending and bridging currently accepted theories of fundamental interactions, such as the Standard Model of particle physics and the theory of general relativity. This includes research on string theory, quantum gravity, theoretical cosmology, and black hole theory. These are all fields that emerged as central after the Standard Model of particle physics became consolidated in the 1980s, and the visual effect may be related to development of string theory.[35] The gray cluster is instead clearly related to particle detectors and instrumentation which, in the context of this research, is most likely related to the rise of what has been called astroparticle physics, namely the field that studies particles from space and their interaction, which emerged as a research area in the early 2000s, in line with what is shown in the figure.[36]

After this iteration, it is clear that the methodology, while still unable to completely disentangle some broader research areas from those strictly connected to

---

33 On the role of the Cold War in the rise of space sciences, see, DeVorkin 1992.
34 Weart 2003.
35 Rickles 2014.
36 Cirkel-Bartelt 2008, p. 2

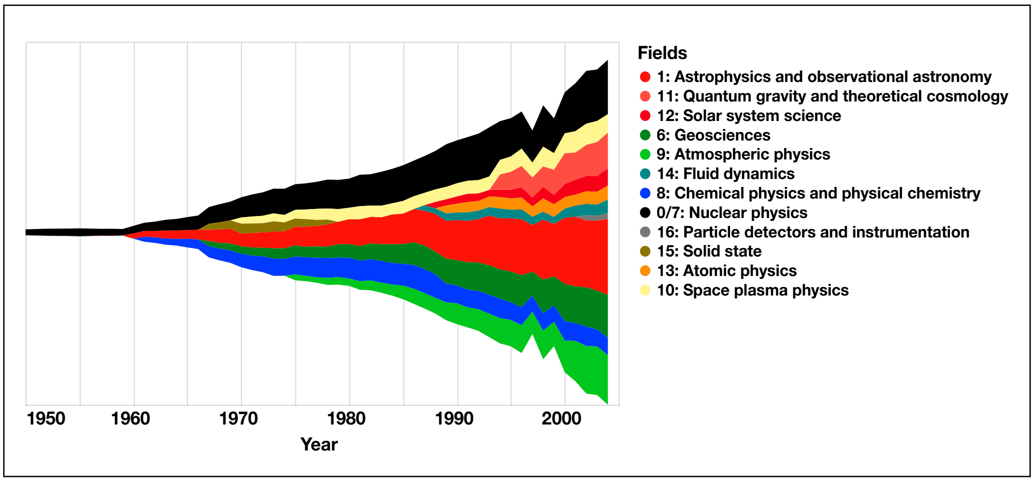

Fig. 7 Streamgraph of cluster size per year for the reduced co-citation network of the ADS corpus (1950–2004) after the second iteration. Data selected for clusters with more than 6000 publications. Field names are based on the interpretation of cluster's content represented in the reports by RL.

astro cluster sciences, provides a picture that aligns quite closely with the broad historical processes that historians have described using more traditional, qualitative approaches.

Temporal evolution of centrality measures

To address the question of the influence of single institutions like the Max Planck Society, we applied a second analytical step to the resulting cluster data. For this approach we considered different networks' centrality measures and compared their values across time. The aim of this analytical step is to allow historians not only to interpret the retrieved clusters in their overall content but also to identify crucial actors in dynamical transformations within a given cluster. To further disentangle the clusters analyzed in the previous section and visible in Fig. 7, the workflow was iterated 3 more times. After each round, a number of clusters were selected to be excluded from the next round and the resolution parameter was adapted. Once a useful set of clusters was found we continued with the analytic workflow.

For the resulting dataset, we calculated the selected normalized centrality for the full co-citation graph of a given year and collected the results in a histogram with logarithmic binning. The value of a given bin is divided by the number of nodes in the given year. This gives the overall yearly distribution of centrality measures for all nodes. Subsequently, nodes from the selected cluster are identified in the calculated centrality values. Finally, a third selection is obtained by identifying publications in the cluster with affiliations of the organization under investigation, which in our case study is the Max Planck Society. Since this identification was already part of the initial data gathering (Fig. 1), we can directly identify these nodes.

Having run these steps for a given year, we must repeat the process for all years within the timeframe of the analysis. Using the normalized approach for all selected centrality measures allows a comparison of the histograms across the vast variety of publications and citations' outputs for the years between 1950 and 2004. The individual datasets are collected in CSV files. For the historical interpretation of the results, the CSV files are read into Pandas DataFrames and plotted in 3D, as detailed below.

For a given centrality measure, three datasets are taken into consideration for each visualization: the overall histogram, the dataset concerning cluster nodes, and the dataset of institutional nodes. To show the position of the cluster and institution nodes, each data point for either the cluster or the institution is added to the value of the respective binned data of the full histogram. The overall centrality measure is visualized as a 3D surface plot, with nodes belonging to the selected cluster shown as black dots, and nodes belonging to the targeted organization as red diamonds. This visualization allows us to compare the influence of the selected cluster on the overall dynamic of the centrality measure, thus highlighting the role of the selected organization in contributing to the centrality as well as the specific point in time.

Fig. 8 shows an example of this visualization for the betweenness centrality measures of a cluster obtained after 5 iterations of exclusion steps with time clustering run with the resolution parameter $7 * 10^{-6}$ and an interslice coupling of 0.3, which was identified as high-energy astrophysics. The diagram shows an overall high degree of involvement for the Max Planck Society in the cluster, as well as its betweenness centrality measures, which can be interpreted as the role of the publications by the Max Planck Society's authors as brokers between different research strands within this large cluster. However, while the 3D plot reveals the existence of the cluster from the start of the research period in 1950 (black dots), the Max Planck Society only appears in the cluster from around 1980, when it started to play a substantial role in connecting it (red diamonds).

The routines for this second analytic step are automated to generate visualizations for all selected clusters and the four centrality measures: authority, betweenness, degree and closeness centrality. With this step we have completed the workflow, from a large, unsorted corpus of unclassified content to a structured view of the research fields related to the chosen case study, which allows us to analyze the involvement of specific institutions in its genesis and historical developments.

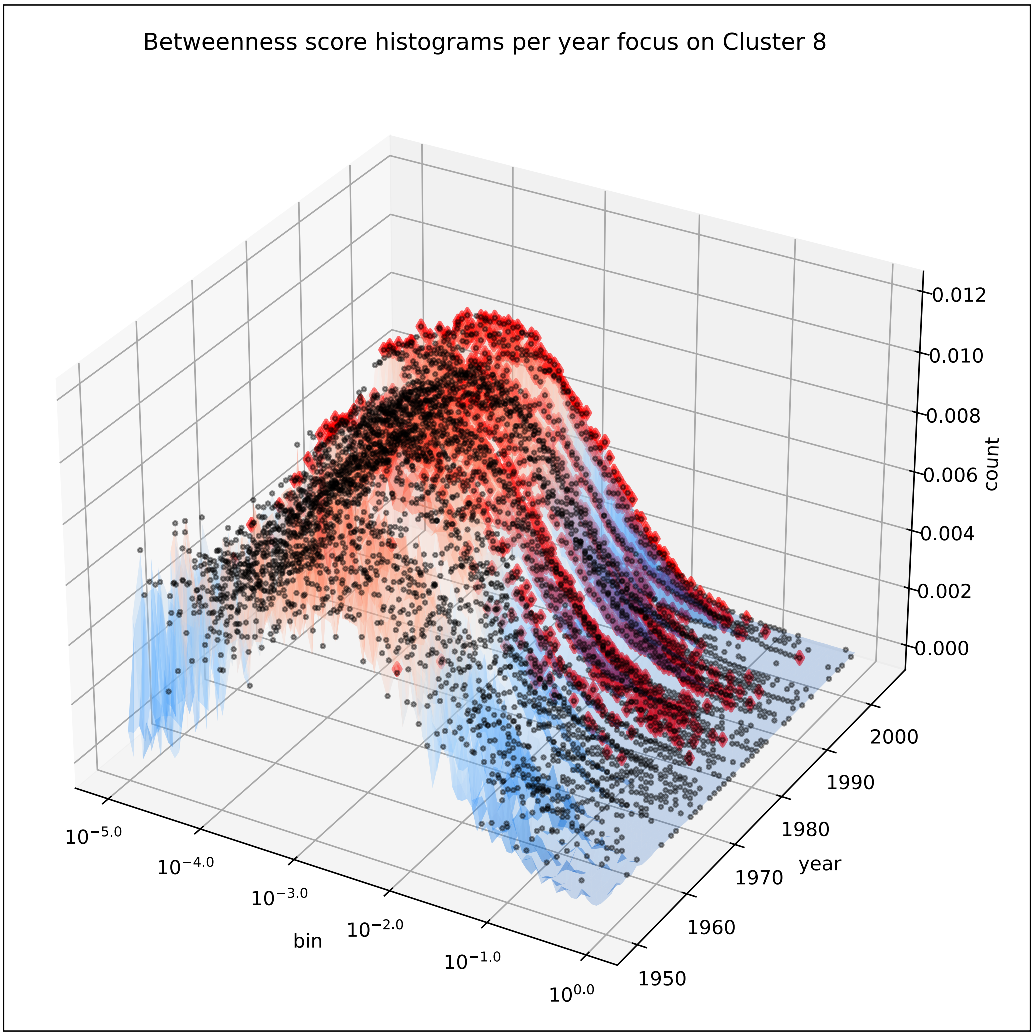

Fig. 8 Normalized betweenness centrality histograms with logarithmic binning (10^{-5} – 10^{0}) for the full co-citation network (surface plot), publications in cluster high-energy astrophysics (black dots), and publications in that cluster with Max Planck Society affiliations (red diamonds) for the years 1950 to 2004. The z-axis shows the relative number of publications in the given bin.

Conclusion and Outlook

The general methodology of the workflow proposed leads to meaningful clusters of three main types, depending on the selected resolution parameter. Some large clusters (approximately larger than 150k publications) clearly group publications concerning broad scientific fields, with many sub-specializations still part of the main cluster. This is the case for clusters related to, for example, solid state physics, quantum optics, superconducting physics and earth-system sciences. In intermediate clusters (in the order of a few 10k publications) we find more methodology-oriented clusters, e.g., on the topic of nuclear magnetic resonance or quantum dot research and technological development. Here, the initial cluster emerges with the introduction of the new technology and grows with later applications of it. A third type of even smaller cluster (less than 5k publications) seems to represent specific collaborations of persons or institutions around shared research themes.

In each iteration, computing time and RAM usage are significantly decreased due to the controlled exclusion resulting from the decisions of the domain expert concerning non-astro related research fields. This allows for the creation of a clean corpus in a reasonable amount of time. In the later steps of the application of the workflow, a new parameter sweep concerning the clustering resolution parameter was necessary to separate the diverse subfields still contained within the largest cluster related to astro cluster sciences.

The corpus thus obtained with the already retrieved affiliation data allows us to return to the original research objective: for each subfield related to astro cluster sciences, the relative importance of publications by scientists from MPG institutes can be calculated separately using different centrality measures. The visualization output then reveals the structural role that the selected institution has had on the entire field. By using the case study of the Max Planck Society, we could then compare the results with those obtained through traditional approaches, usually finding supporting evidence.[37] In addition to this, the cleaned corpora of specific subfields also allow us to address questions on the role of general historical and political events.[38]

In future work, we will address the evolution of a wider range of subfields in physics, such as material and laser science, compare our results with previous analysis,[39] and collaborate with experts in the history of medicine and earth-system science to further elaborate the usability of the workflow for other, more interdisciplinary research fields.

Another line of ongoing research concerns the application of large language models for the identification of cluster content, which in particular simplifies the search for suitable parameter settings for a given research question. Nonetheless, a full interpretation of the historical settings involved in the evolution of a scientific field remains firmly in the capable hands of historians.

---

37 Bonolis and Leon 2022.
38 Borjas and Doran 2014.
39 For instance, Sinatra et al. 2015.

Bibliography

Aronova, Elena. "Geophysical Datascapes of the Cold War: Politics and Practices of the World Data Centers in the 1950s and 1960s." *Osiris* 32, no. 1 (September 2017): 307–27. https://doi.org/10.1086/694094.

Bettencourt, Luís M. A., David I. Kaiser, Jasleen Kaur, Carlos Castillo-Chávez, and David E. Wojick. "Population Modeling of the Emergence and Development of Scientific Fields." *Scientometrics* 75, no. 3 (June 1, 2008): 495. https://doi.org/10.1007/s11192-007-1888-4.

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Katrin Menzel, and Elke Teich. "The Diffusion of Scientific Terms – Tracing Individuals' Influence in the History of Science for English." In *Proceedings of the 5th Joint Sighum Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature,* edited by Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, 120–27. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, 2021. doi: https://doi.org/10.18653/v1/2021.latechclfl-1.14.

Bonolis, Luisa, and Juan-Andres Leon. *Astrophysics, Astronomy and Space Sciences in the History of the Max Planck Society.* Brill, 2022. https://brill.com/display/title/59720.

Bonolis, Luisa, and Juan-Andres Leon. "Gravitational-Wave Research as an Emerging Field in the Max Planck Society: The Long Roots of Geo600 and of the Albert Einstein Institute." In *The Renaissance of General Relativity in Context,* edited by Alexander S. Blum, Roberto Lalli, and Jürgen Renn, 285–361. Cham: Springer International Publishing, 2020. doi: https://doi.org/10.1007/978-3-030-50754-1_9.

Borjas, George, and Kirk Doran. "Which Peers Matter? The Relative Impacts of Collaborators, Colleagues, and Competitors." Cambridge, MA: National Bureau of Economic Research, March 2014. doi: https://doi.org/10.3386/w20026.

Bornmann, Lutz. "Field Classification of Publications in Dimensions: A First Case Study Testing Its Reliability and Validity." *Scientometrics* 117, no. 1 (October 2018): 637–40. doi: https://doi.org/10.1007/s11192-018-2855-y.

Boyack, Kevin W., and Richard Klavans. "Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?" *Journal of the American Society for Information Science and Technology* 61, no. 12 (December 1, 2010): 2389–2404. https://doi.org/10.1002/asi.21419.

Chen, Chaomei, and Min Song. "Visualizing a Field of Research: A Methodology of Systematic Scientometric Reviews." *PLOS ONE* 14, no. 10 (October 31, 2019): e0223994. https://doi.org/10.1371/journal.pone.0223994.

Cirkel-Bartelt, Vanessa. "History of Astroparticle Physics and Its Components." *Living Reviews in Relativity* 11, no. 1 (May 7, 2008): 2. https://doi.org/10.12942/lrr-2008-2.

DeVorkin, David H. *Science with a Vengeance: How the Military Created the US Space Sciences after World War II.* New York: Springer-Verlag, 1992.

DeVorkin, David, The Space Age and Disciplinary Change in Astronomy. In: Steven J. Dick (ed.): *NASA's First 50 Years. Historical Perspectives.* Washington, D.C.:NASA, 2010, 389–426.

DeWilde, Burton, and Contributors. "Textacy." URL: https://github.com/chartbeat-labs/textacy (last viewed June 5th, 2024).

Févotte, Cédric, and Jérôme Idier. "Algorithms for Nonnegative Matrix Factorization with the β-Divergence." *Neural Computation* 23, no. 9 (September 2011): 2421–56. doi: https://doi.org/10.1162/NECO_a_00168.

Galison, Peter, and Bruce William Hevly. *Big Science: The Growth of Large-Scale Research.* Stanford, Calif.: Stanford University Press, 1992.

Harzing, Anne-Wil. "Two New Kids on the Block: How Do Crossref and Dimensions Compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?" *Scientometrics* 120, no. 1 (July 2019): 341–49. doi: https://doi.org/10.1007/s11192-019-03114-y.

Kaye, Aleksandra, Raphael Schlattmann, Malte Vogl, Bernardo S. Buarque, Jascha Schmitz, Lea Weiß and Laura von Welczeck. "Socio-Epistemic Networks: a Framework for History of Knowledge" in *History of Intellectual Culture 3/2024: Experimental Spaces: Knowledge Production and its Environments in the Long Nineteenth Century,* edited by Charlotte A. Lerg, Johan Östling, Jana Weiß, Anne Kwaschik, and Claudia Roesch. Oldenbourg: De Gruyter, 2024. doi: https://doi.org/10.1515/9783111291383-012.

Klavans, Richard, and Kevin W. Boyack. "Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge?" *Journal of the Association for Information Science and Technology* 68, no. 4 (2017): 984–98. https://doi.org/10.1002/asi.23734.

Lalli, Roberto, Riaz Howey, and Dirk Wintergrün. "The Dynamics of Collaboration Networks and the History of General Relativity, 1925–1970." *Scientometrics* 122, no. 2 (February 2020b): 1129–70. doi: https://doi.org/10.1007/s11192-019-03327-1.

—. "The Socio-Epistemic Networks of General Relativity, 1925–1970." In *The Renaissance of General Relativity in Context,* edited by Alexander S. Blum, Roberto Lalli, and Jürgen Renn, 15–84. Einstein Studies. Cham: Springer International Publishing, 2020a. doi: https://doi.org/10.1007/978-3-030-50754-1_2.

Lancichinetti, Andrea, and Santo Fortunato. "Community Detection Algorithms: A Comparative Analysis." *Phys. Rev. E* 80, no. 5 (November 2009): 056117. doi: https://doi.org/10.1103/PhysRevE.80.056117.

Martin, Joseph D. "When Condensed-Matter Physics Became King." *Physics Today* 72, no. 1 (January 1, 2019): 30–37. https://doi.org/10.1063/PT.3.4110.

Mucha, Peter J., Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. "Community Structure in Time-Dependent, Multiscale, and Multiplex Networks." *Science* 328, no. 5980 (2010): 876–78. doi: https://doi.org/10.1126/science.1184819.

Perianes-Rodriguez, Antonio, and Javier Ruiz-Castillo. "A Comparison of the Web of Science and Publication-Level Classification Systems of Science." *Journal of Informetrics* 11, no. 1 (February 1, 2017): 32–45. https://doi.org/10.1016/j.joi.2016.10.007.

Reinhardt, Carsten. III. 2 Strukturen der Forschung, in *Die Max-Planck-Gesellschaft: Wissenschafts- Und Zeitgeschichte 1945–2005,* ed. Jürgen Renn, Carsten Reinhardt, and Jürgen Kocka (Göttingen: Vandenhoeck & Ruprecht, 2024), p. 224, doi: https://doi.org/10.13109/9783666993640.203.

Renn, Jürgen, Carsten Reinhardt, and Jürgen Kocka (ed.). *Die Max-Planck-Gesellschaft: Wissenschafts- und Zeitgeschichte 1945–2005.* Göttingen: Vandenhoeck & Ruprecht, 2024, doi: https://doi.org/10.13109/9783666993640.

Renn, Jürgen, Dirk Wintergrün, Roberto Lalli, Manfred Dietrich Laubichler, and Matteo Valleriani. "Netzwerke als Wissensspeicher." In *Die Zukunft der Wissensspeicher: Forschen, Sammeln und Vermitteln im 21. Jahrhundert,* edited by Jürgen Mittelstraß and Ulrich Rüdiger, 7:35–79. Konstanzer Wissenschaftsforum. München: UVK Verlagsgesellschaft Konstanz, 2016.

Rickles, Dean. *A Brief History of String Theory: From Dual Models to M-Theory.* Berlin, Heidelberg: Springer Berlin Heidelberg: Imprint: Springer, 2014.

Sinatra, Roberta, Pierre Deville, Michael Szell, Dashun Wang, and Albert-László Barabási. "A Century of Physics." *Nature Physics* 11, no. 10 (October 2015): 791–96. doi: https://doi.org/10.1038/nphys3494.

Traag, V. A., L. Waltman, and N. J. van Eck. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports* 9, no. 1 (March 2019): 5233. doi: https://doi.org/10.1038/s41598-019-41695-z.

Traag, V., and Contributors. "Leidenalg.". URL: https://github.com/vtraag/leidenalg (last viewed June 5th, 2024).

Traag, V. A., P. Van Dooren, and Y. Nesterov. "Narrow Scope for Resolution-Limit-Free Community Detection." *Physical Review E* 84, no. 1 (July 29, 2011): 016114. https://doi.org/10.1103/PhysRevE.84.016114.

Turchetti, Simone, and Peder Roberts, eds. *The Surveillance Imperative: Geosciences during the Cold War and Beyond.* New York: Palgrave Macmillan US, 2014. https://doi.org/10.1057/9781137438744.

Vanderplas, Jake, and Contributors. "Altair." URL: https://github.com/altair-viz/altair (last viewed June 5th, 2024).

Wagner, Caroline S., and Loet Leydesdorff. "Seismology as a Dynamic, Distributed Area of Scientific Research." Scientometrics 58, no. 1 (September 1, 2003): 91–114. https://doi.org/10.1023/A:1025479524390.

Weart, Spencer R. *The Discovery of Global Warming.* New Histories of Science, Technology, and Medicine. Cambridge, Mass.: Harvard University Press, 2003.